Introduction to Statistical Methods in Political Science

Lecture 9: Inference for Two Proportions

Ignacio Urbina

Ph.D. Candidate in Political Science

Motivation: Inference for Categorical Differences

- Categorical variables represent data sorted into distinct groups or categories.
- Often, we are interested in comparing proportions of a categorical outcome between two groups.
- Example: In a survey, respondents might be grouped by region to see if a higher proportion of people in one region favor a particular policy option compared to another.
- These comparisons help answer questions such as:
  - Does the proportion of people supporting stricter environmental regulations differ between urban and rural areas?
  - Is there a difference in preferred government spending options between younger and older age groups?

Introduction to Comparing Proportions

- In statistics, we often compare two independent groups to understand differences in proportions.
- Examples include comparing vaccination rates between cities or customer satisfaction across products.
- Proportions are used in medicine, marketing, public health, and social sciences.
- These comparisons help identify significant differences in behaviors, treatments, or characteristics.

Example Scenarios

- Medical studies: Comparing success rates of medications.
- Marketing: Evaluating customer satisfaction between products.
- Social sciences: Analyzing behavior differences between demographic groups.

Example: Public Opinion

One common application of comparing two sample proportions is in analyzing public opinion across different demographic groups:

- Consider two groups: those under 30 and those over 60.
- Survey both groups to assess their support for the current president.
- Let $\hat{p}_1$ be the proportion of support among those under 30 and $\hat{p}_2$ among those over 60.

This analysis helps us understand how support varies with age, crucial for policy making and election strategies.

Example: Psychological Experiment

In psychological research, comparing sample proportions evaluates the effect of different treatments:

- Study testing two interventions on anxiety reduction.
- One group receives cognitive behavioral therapy (CBT); another receives mindfulness-based stress reduction (MBSR).
- Let $\hat{p}_1$ be the proportion reporting significant anxiety reduction with CBT, and $\hat{p}_2$ with MBSR.

This comparison provides insights into which intervention is more effective.

Notation and Definitions

- Let p_1 and p_2 represent the population proportions for two independent groups.
- $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions from each group.
- Sample sizes: n_1 for Group 1 and n_2 for Group 2.

Introduction to Sample Proportions

Consider two independent samples where:

- Sample 1: n_1 observations with proportion $\hat{p}_1$.
- Sample 2: n_2 observations with proportion $\hat{p}_2$.

We are interested in the statistic $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions.

- Our goal is to estimate the difference $p_1 - p_2$ and determine if it is significantly different from zero.
- This analysis helps assess if observed differences could be due to chance or represent a true population difference.

Review of Expectations and Variances

Expectations:

- For random variables X and Y, and constants a, b:

$$E(aX + bY) = aE(X) + bE(Y)$$

Variances:

- For independent random variables X and Y, and constants a, b:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

Expectation of the Statistic

Using linearity of expectations:

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

Where p_1 and p_2 are the true population proportions.

Variance of the Statistic

For independent random variables:

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2)$$

Given:

$$\mathrm{Var}(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}, \quad \mathrm{Var}(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

Thus:

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Standard Error of the Statistic

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\mathrm{Var}(\hat{p}_1 - \hat{p}_2)}$$

Substituting the variances:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

Sampling Distribution of Difference in Proportions

- The distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal for large sample sizes.
- Under the Central Limit Theorem:

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- This allows us to perform hypothesis testing and construct confidence intervals.

Derivation of the Sampling Distribution

- Variance of a sample proportion $\hat{p}$: $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$
- Variance of the difference:

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- Standard error (SE):

$$SE = \sqrt{\text{Var}(\hat{p}_1 - \hat{p}_2)}$$

Confidence Interval for Difference in Proportions

- Confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE$$

- Where:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- z^* is the critical value for the desired confidence level (e.g., 1.96 for 95% confidence).

Constructing the Confidence Interval

1. Calculate sample proportions: $\hat{p}_1$ and $\hat{p}_2$.
2. Compute the standard error (SE):

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

3. Determine critical value z^* for the desired confidence level.
4. Multiply z^* by SE to find the margin of error (MOE).
5. Construct the confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm MOE$$

Applied Example: Comparing Vaccination Rates - Setup

- Sample data:
  - Group 1: $n_1 = 400$, $\hat{p}_1 = 0.30$
  - Group 2: $n_2 = 500$, $\hat{p}_2 = 0.25$
- We will calculate the standard error and confidence interval.

Applied Example: Comparing Vaccination Rates - Calculation

- Calculate SE:

$$SE = \sqrt{\frac{0.30 \times 0.70}{400} + \frac{0.25 \times 0.75}{500}} \approx 0.032$$

- 95% Confidence Interval:

$$(0.30 - 0.25) \pm 1.96 \times 0.032$$

$$0.05 \pm 0.063 \implies [-0.013, 0.113]$$

Interpretation of Example Results

- Confidence interval includes zero (-0.013 to 0.113), suggesting no significant difference.
- If the interval excluded zero, it would indicate a statistically significant difference.

Common Misconceptions

- **Misconception 1**: A confidence interval that includes zero proves there is no difference between the two population proportions.
  - **Reality**: It simply suggests we lack sufficient evidence of a difference at the specified confidence level.
- **Misconception 2**: A wider confidence interval means the difference is less likely.
  - **Reality**: A wider interval indicates greater uncertainty, often due to smaller sample sizes.
- **Misconception 3**: A 95% confidence level means the interval has a 95% chance of containing the true difference.
  - **Reality**: A 95% confidence level means that, in the long run, 95% of intervals from multiple independent samples of the same size will contain the true difference.

Summary and Key Takeaways

- Inference for difference in proportions helps compare two independent groups.
- Large samples allow the sampling distribution to be approximately normal.
- Confidence intervals offer a range of plausible values for the true difference.
- The examples illustrated the application and interpretation of these concepts.

Hypothesis Testing for Difference in Proportions

- Beyond estimating the confidence interval, we often want to test whether the observed difference between two proportions is statistically significant.
- Hypothesis testing allows us to determine if there is enough evidence to support a claim about the difference in population proportions.
- Common applications include testing the effectiveness of a new treatment compared to a control or comparing preferences between two groups.

Formulating Hypotheses

- **Null Hypothesis (H_0)**: Assumes no difference between the population proportions.

$$H_0 : p_1 - p_2 = 0$$

- **Alternative Hypothesis (H_a)**: Proposes that there is a difference.

$$\text{Two-tailed test: } H_a : p_1 - p_2 \neq 0$$

$$\text{One-tailed test: } H_a : p_1 - p_2 > 0 \text{ or } p_1 - p_2 < 0$$

- The choice between one-tailed and two-tailed tests depends on the research question.

Test Statistic

- The test statistic measures how far the sample difference is from the null hypothesis, relative to the standard error.
- Under H_0, the test statistic is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_0}$$

- Where SE_0 is the standard error calculated under the assumption that H_0 is true.

Standard Error Under Null Hypothesis

- Under $H_0 : p_1 = p_2 = p$, we pool the sample proportions to estimate the common population proportion p.
- The pooled sample proportion $\hat{p}$ is:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- Where x_1 and x_2 are the number of successes in each sample.
- The standard error under H_0 is:

$$SE_0 = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Calculating the Test Statistic

1. Compute the pooled sample proportion $\hat{p}$:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

2. Calculate the standard error SE_0 under H_0:

$$SE_0 = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

3. Compute the test statistic z:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{SE_0}$$

Decision Rule and p-value

- **Decision Rule**:
  - Compare the test statistic z to critical values from the standard normal distribution.
  - For a two-tailed test at $\alpha = 0.05$, reject H_0 if $|z| > 1.96$.
- **p-value**:
  - The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated, assuming H_0 is true.
  - For a two-tailed test:

$$p\text{-value} = 2 \times P(Z > |z|)$$

- **Conclusion**:
  - If the p-value is less than α, reject H_0.
  - Otherwise, fail to reject H_0.

Applied Example: Hypothesis Testing - Setup

- Continuing the vaccination rate example:
  - Group 1: $n_1 = 400$, $\hat{p}_1 = 0.30$, $x_1 = 120$
  - Group 2: $n_2 = 500$, $\hat{p}_2 = 0.25$, $x_2 = 125$
- Test whether there is a significant difference in vaccination rates between the two groups at the 5% significance level.
- Formulate hypotheses:

$$H_0 : p_1 - p_2 = 0 \quad \text{vs.} \quad H_a : p_1 - p_2 \neq 0$$

Applied Example: Hypothesis Testing - Calculations

1. Compute pooled proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 125}{400 + 500} = \frac{245}{900} \approx 0.272$$

2. Calculate standard error under H_0:

$$SE_0 = \sqrt{0.272 \times 0.728 \left(\frac{1}{400} + \frac{1}{500} \right)}$$
$$\approx \sqrt{0.198 \times 0.0045} \approx 0.0299$$

3. Compute test statistic:

$$z = \frac{0.30 - 0.25}{0.0299} \approx \frac{0.05}{0.0299} \approx 1.675$$

Applied Example: Decision and Conclusion

- Critical value for two-tailed test at $\alpha = 0.05$ is $z_{\alpha/2} = 1.96$.
- Since $|z| = 1.675 < 1.96$, we fail to reject H_0.
- **p-value**:

$$\begin{aligned} p\text{-value} &= 2 \times P(Z > 1.675) \\ &= 2 \times (1 - \Phi(1.675)) \\ &\approx 2 \times 0.04697 = 0.094 \end{aligned}$$

- Since p-value $= 0.094 > 0.05$, we fail to reject H_0.
- **Conclusion**:
  - There is insufficient evidence at the 5% significance level to conclude a difference in vaccination rates between the two groups.

Interpretation of Example Results

- Although the sample proportions differ (30% vs. 25%), the difference is not statistically significant at the 5% level.
- This suggests that the observed difference could be due to random sampling variability.
- It's important to consider sample sizes and variability when interpreting results.

Common Misconceptions

- **Misconception 1**: A non-significant result means there is no difference.
  - **Reality**: It means we do not have sufficient evidence to conclude a difference exists.
- **Misconception 2**: A small p-value indicates a large effect size.
  - **Reality**: The p-value measures evidence against H_0, not the magnitude of the effect.
- **Misconception 3**: Failing to reject H_0 proves H_0 is true.
  - **Reality**: We can never prove H_0; we can only fail to reject it.

Summary and Key Takeaways

- Hypothesis testing for the difference in proportions assesses whether an observed difference is statistically significant.
- The test involves calculating a test statistic under the assumption that the null hypothesis is true.
- Understanding the standard error and using the pooled proportion are critical steps.
- The p-value helps determine the strength of evidence against the null hypothesis.
- Always interpret results in context and be cautious of common misconceptions.